Ohio J. Sch. Math. No. 10, Jan. 1986

THE GEOMETRIC PROBABILITY DISTRIBUTION: A LESSON FOR COLLECTORS

David Kullman and John Skillings
Miami University, Oxford, Ohio

What do the following problems have in common?

Cereal Box Problem: Each box of a certain breakfast cereal contains a prize. If there are six different prizes in the set, how many boxes of cereal should a person have to buy in order to collect the entire set of prizes?

State Flags Problem: A few years ago the United States Postal Service issued a set of 50 commemorative stamps featuring state flags. (Later a second set of stamps, showing state birds and flowers, was issued.) If a stamp collector attempts to obtain all 50 state flag stamps by examining old envelopes, how many envelopes with flag stamps should he or she expect to examine before completing the collection?

Baseball Card Problem: Baseball cards are randomly packaged and sold. If there are 750 different baseball cards in a complete year's set, how many cards should a collector have to purchase (without being able to select individual cards) in order to obtain every card?

Each of these problems can be solved by an application of the geometric probability distribution. This is so-named because it is computed from the partial sums of a geometric series. While several articles have suggested ways of simulating these problems using Monte-Carlo methods, we will present a theoretical solution suitable for high school students of probability. Before stating a general result, we will illustrate the probabilities involved in the cereal box problem.

Look at the flow graph for the cereal box problem in Figure 1 at the end of this article. Each circle represents a state corresponding to the number of different prizes that have already been obtained. Each arrow represents a trial that consists of examining the prize in a new box of cereal. If that prize is different from the others, we advance to the next state. Otherwise, we remain in the same state. The number next to each arrow represents the probability of that outcome.

Four assumptions are made: (1) the prizes are randomly distributed among the boxes; (2) each trial results in a success (getting a new prize) or failure (getting a duplicate of a previous prize); (3) the trials are repeated until a success occurs; (4) the trials are independent (i.e., the outcome on one trial does not affect the outcome on any other trial). The solution to our problem (how many boxes should be opened to collect a complete set of prizes) will be the expected number of trials needed to move from the initial state (no prizes) to the final state (six prizes). By expected number, we mean an average of the number of trials needed if the experiment were repeated many times. In solving the problem by Monte Carlo methods, a simulated experiment is, in fact, repeated over and over, and the average number of trials is computed.

Our experiment is actually a sequence of six simpler experiments, each consisting of a move from one state to the next. Consequently, the expected number of trials needed to collect all six prizes will be the sum of the expected number of trials needed to obtain each prize. Suppose we have already found two different prizes and are looking for the third. The expected number of trials is, by definition, the sum of terms of the form $x \cdot f(x)$, where x is a number of trials and $f(x)$ is the probability that exactly x trials will be required. In particular, when moving from state 2 to state 3

$f(x)$ = P (x trials are needed to move from state 2 to state 3)
= P (first x - 1 trials are all failures and the last trial is a success)
$= (\frac{2}{6})^{x-1} (\frac{4}{6})$.

Using this for $f(x)$ we find that the expected number of trials needed to move from state 2 to state 3 is

$$1 \cdot (\tfrac{4}{6}) + 2 \cdot (\tfrac{2}{6})(\tfrac{4}{6}) + 3(\tfrac{2}{6})^2(\tfrac{4}{6}) + \cdots + x(\tfrac{2}{6})^{x-1}(\tfrac{4}{6}) + \cdots = \sum_{x=1}^{\infty} x(\tfrac{2}{6})^{x-1}(\tfrac{4}{6}).$$

There is no theoretical upper bound to the number of terms, since it is conceivable that we may find nothing but duplicates.

We could write analogous sums for the expected number of trials needed to move to each of the other states, but let's look

at the more general problem now. Let X be the number of trials needed for a success (getting a new prize), p equal the probability of success on any one trial, and q = 1 - p equal the probability of failure on any one trial. The E(X), the expected value of X, can be represented as

$$E(X) = 1 \cdot p + 2 \cdot qp + 3 \cdot q^2p + 4 \cdot q^3p + \cdots + xq^{x-1}p + \cdots = \sum_{x=1}^{\infty} xq^{x-1}p$$

It is helpful to look at this expression in a triangular form:

$$\begin{aligned} E(X) = {} & p + qp + q^2p + q^3p + \cdots + q^{x-1}p + \cdots \\ & + qp + q^2p + q^3p + \cdots + q^{x-1}p + \cdots \\ & + q^2p + q^3p + \cdots + q^{x-1}p + \cdots \\ & + q^3p + \cdots + q^{x-1}p + \cdots \\ & \text{etc.} \end{aligned}$$

Factoring, E(X) may be written as

$$\begin{aligned} E(X) = {} & (1 + q + q^2 + q^3 + \cdots + q^{x-1} + \cdots)\, p \\ & + (1 + q + q^2 + q^3 + \cdots + q^{x-1} + \cdots)\, qp \\ & + (1 + q + q^2 + q^3 + \cdots + q^{x-1} + \cdots)\, q^2p \\ & + (1 + q + q^2 + q^3 + \cdots + q^{x-1} + \cdots)\, q^3p \\ & \text{etc.} \end{aligned}$$

Note that each expression in parentheses is the same geometric series, $\sum_{x=0}^{\infty} q^x$, which sums to $\frac{1}{1-q}$, while the column of terms at the right sums to

$$p \sum_{x=0}^{\infty} q^x = \frac{p}{1-q}. \quad \text{Thus } E(X) = \frac{p}{(1-q)^2}, \text{ but } p = 1 - q, \text{ so}$$

$$E(X) = \frac{p}{p^2} = \frac{1}{p}.$$

What could be simpler? But remember, this is only part of our answer. The expected <u>total</u> number of trials will be the sum of terms $1/p_i$, where p_i is the probability of a success on each trial as we seek to obtain the <u>i</u>th distinct prize. In particular, for the cereal box problem with six prizes, we expect to open

$$E = \frac{1}{1} + \frac{1}{5/6} + \frac{1}{4/6} + \frac{1}{3/6} + \frac{1}{2/6} + \frac{1}{1/6}$$

$$= 1 + 6/5 + 3/2 + 2 + 3 + 6 = 14.7 \text{ boxes.}$$

In other words, if many persons try to collect the set of six prizes, the <u>average</u> number of boxes of cereal that each will have to buy should be about 15. This is two and one-half times the number of prizes involved. Of course, we are assuming that the prizes are randomly distributed among cereal boxes, and that no trading occurs between collectors. Clearly, the model will not be valid if only five of the six prizes are sent to our town, and it should be equally clear that it will be advantageous for collectors to get together to trade their duplicates.

What about the related stamp collecting and baseball card problems, in which larger numbers of "prizes" are involved? For n prizes, the values of p_i are $1 = \frac{n}{n}, \frac{n-1}{n}, \frac{n-2}{n}, \frac{n-3}{n}, \ldots, \frac{1}{n}$, respectively. Consequently, the expected number of trials is

$$E = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \frac{n}{n-3} + \ldots + \frac{n}{1}$$

$$= n\left(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \frac{1}{n-3} + \ldots + \frac{1}{1}\right)$$

$$= n \sum_{k=1}^{n} 1/k.$$

To solve this we need to compute partial sums of the harmonic series $\sum_{k=1}^{n} 1/k$. This may be accomplished by summing the terms with the aid of a computer or hand-held calculator, but there is an elegant approximation that makes use of Euler's constant, γ, which is approximately 0.5772. The approximation is $\sum_{k=1}^{n} 1/k \approx \log_e n + \gamma + 1/(2n)$. Multiplying by n gives the expected number of trials.

For the cereal box problem, this approximation yields $6(\log_e 6 + \gamma + 1/12) = 6(1.792 + .577 + .083) = 14.7$, agreeing with our earlier result. The same approximation method shows that a stamp collector should expect to examine about 225 envelopes with flag stamps before obtaining a complete set of 50. For a set of 750 baseball cards, the collector can expect to go through nearly 5,400 before obtaining those elusive last few.

Notice that the ratio of expected trials to prizes increases with n, the number of different prizes. As n gets large, this ratio approaches $\log_e n$. So the "lone wolf" collector is at a distinct disadvantage, compared to those who trade with each other. This probably explains the prolific growth of organizations for just about every kind of collectible item imaginable.

REFERENCES

Lappan, Glenda and Winter, M.J. Probability Simulation in Middle School. Mathematics Teacher, 73 (1980), 446-449.

Mosteller, Frederick. Collecting Coupons. Problem 14 in Fifty Challenging Problems in Probability, Addison Wesley, 1965.

Shiflett, Ray C. and Shultz, Harris S. Can I Expect a Full Set? Mathematical Gazette, 64 (1980), 262-266.

Mind Stretchers

1. An absent-minded professor cashed a check and was inadvertently paid in dollars what the check called for in cents, and in cents what was called for in dollars. He did not notice this until later, when he found that he had twice as much money in his pocket as he should have had. By that time he had spent 41¢ . How much did he have left?

2. Five women, each accompanied by one daughter, are buying cloth in a shop. Every one of the ten buys as much cloth in yards as she pays cents per yard for her purchases. Every mother spends $4.05 more than her daughter. Mrs. Evans buys 23 yards less than Mary; Rose spends 9 times as much as Clara, and Effie buys 8 yards less than 10 times as much as Mrs. Jones. Mrs. Connor spends $40.32 more than Mrs. Smith, but Mrs. Brown spends most of all. Now, what is Helen's last name ?

3. There are many sets of positive integers which can be formed by using each of the ten digits exactly once. For example, 347, 256, 810, and 9 form such a set. Also, 12, 341 56, 78, and 90 form such a set. Can you find such a set of positive integers for which the sum of the numbers in the set is 100 ? Either find such an example or explain why none exists.